STIC-ILL

105/23

From: Zeman, Mary
Sent: Friday, May 23, 2003 1:26 AM
To: STIC-ILL
Subject: References 09/96105

447315

Please send a copy of the following references
Thank you
Mary K. Zeman
Examiner, 1631
305-7133
CM1 12A17
mailbox: CM1 12D01

SO Methods in Molecular Biology (Totowa, New Jersey) (2000),
132(Bioinformatics Methods and Protocols) 71.91
CODEN: MMBIED; ISSN: 1064.3745

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 1.7 ISSN: 1415.

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 9.15 ISSN: 1415.4757

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 17.23 ISSN: 1415.4757

SO International Genome Sequencing and Analysis Conference, (2000) Vol. 12,
pp. 97.98. print.
Meeting Info.: 12th International Genome Sequencing and Analysis
Conference Miami Beach, Florida, USA September 12.15, 2000

4540.525000

BRI 5/27

known sequence contexts for the purpose of SNP screening. The reaction is designed as a premix that contains all of the components except primers and templates. The completed reaction identifies one nucleotide located 3' relative to the primer site. We have reformulated our SNaPshot reagent mix to enable robust multiplex SNP interrogation against multiple templates in varying amounts. The resulting multiple products can then be analyzed by electrophoresis in the presence of a size standard, labeled with a 5th dye. Evaluations on ABI Prism Models 310, 377, 3100 and 3700 have been successful. Topics including throughput, reaction format, primer design and template requirements will be covered.

P-178
5-Dye System Compatibility across ABI PRISM® Instrument Platforms

A. Wheaton, D. Wei, C. Holt, S. Menchen, P. Kenny, B. Rosenblum, P. Hanachi, G. Mitra, G. Ayanglou, P. Dong, PE Biosystems, Foster City, CA

We describe the implementation of our new GeneScan™ 5-dye system across all ABI PRISM® instrument platforms. In the 5-dye system the current D dye set (6FAM™, HEX, NED™, and ROX™) becomes the G5 set for higher throughput in major GeneScan applications. G5 includes 6FAM™, VIC™, NED™, PET™, and the new 5th-dye labeled size standard. Likewise, the current E-dye set (dR110, dR6G, dTAMARA™, and dROX™) becomes the E5 set with the addition of the 5th-dye size standard to facilitate automated data analysis. The 5-dye system incorporates 5-dye data collection, optimized running methods, and enhanced analysis tools to enable a variety of new 5-dye applications. Benefits of the 5-dye system include optimal spectral resolution, reliable signal intensity, and increased throughput, across all ABI PRISM® instrument platforms. We display the performance of this new five-dye system across the ABI PRISM® 310 DNA Sequencer, the ABI PRISM® 377, 3700, and newly introduced 3100 DNA Analyzers.

P-179
The Comprehensive Microbial Resource

Owen White, Jeremy Peterson, Jonathan A. Eisen, and Steven L. Salzberg, The Institute for Genomic Research, Rockville, MD

One of the challenges presented by large-scale genome sequencing efforts is the effective display of information in a format that is accessible to the laboratory scientist. Conventional databases offer the scientist the means to search for a particular gene, sequence, or organism, but do little in the way of displaying the vast amounts of curated information that are becoming available. TIGR has developed methods to effectively "slice" the vast amounts of data in the sequencing databases in a wide variety of ways, allowing the user to formulate queries that search for specific genes as well as to investigate broader topics, such as genes that might serve as vaccine and drug targets. The Omniome database contains all of the fully sequenced microbial genomes, the curation from the original sequencing centers, and further curation from TIGR (for those genomes sequenced outside TIGR). The web presentation of the Omniome includes the comprehensive collection of bacterial genome sequences, curated information, and related informatics methodologies. The scientist can view genes within a genome and can also link across to related genes in other genomes. The effect is to be able to construct queries that include sequence searches, isoelectric point, GC-content, GC-skew, functional role assignments, growth conditions, environment and other questions, and isolate the genes of interest. The database contains extensive curated data as well as pre-run homology searches to facilitate data mining. The interface allows the display of the results in numerous formats that will help the user ask more accurate questions. This resource should be of value to the scientific community to design experiments and spur further research. Resources of this type are an essential tool to make sense of bacterial genome information as the number of completed genomes continues to grow.

P-180
How Deep Is Deep Enough: Criteria for planning EST Sequencing Projects

Joseph A. White, Catherine M. Ronning, and John Quackenbush, The Institute for Genomic Research, Rockville, MD

To evaluate the yield of unique sequences obtained from cDNA libraries, sequences were selected at random from known libraries of ESTs. After assembling the sequences into contigs, the numbers of contigs and singletons were counted. We observed the following: 1) the percentages of unique sequences and singleton sequences always decline as sample size is increased, 2) the number of contigs increased linearly over the range of samples sizes selected for this study, 3) the number of singletons approaches a plateau as the sample size is increased, 4) different cDNA libraries have significantly different numbers of unique sequences for the same sample size, and 5) pooling samples from different cDNA libraries increases the number and percentage of unique sequences for the same sample size. Unique sequences are defined as the number of contigs plus the number of singletons. Although this measure of unique sequences is commonly used, it is affected greatly by the number of singletons, which was observed to vary with library and sample size. Although useful, a better measure of uniqueness needs to be obtained.

P-181
Automation of Processes in a Core DNA Sequencing Facility

Suzanne P. Williams[1], Yvette C. Clancy[1], Melissa T. Cronan[1], Kevin J. Laddison[1], Alison E. Maurice[1], David P. Dean[1], Ke Xu[1], Michael Polchaninoff[2], Judith A. Nolan[3], Howard D. Cash[4], Beth A. Clark[4], [1]Pfizer, Groton, CT, [2]Visual Technologies, Portland, CT, [3]PE

Informatics, Foster City, CA, [a]Gene Codes, Ann Arbor, Michigan

Several of the processes in a core DNA Sequencing facility have been optimized to accommodate the increased throughput made possible with the 3700 DNA Analyser. Improvement of several processes will be reported: sample submission, reaction set up, data management and DNA Analysis. Both in-house development and commercial products were considered. The Pfizer DNA Sequencing Facility generates and analyses DNA sequence from samples submitted by scientists site wide. Manual entry of the sample information for the 373 and 3700 Data Collection Software, PE Biosystems, can take up to an hour a day. A web program has been developed for sample submission. The submission information can now be copied and pasted to populate the Collection Sample Sheets. Manual performance of Terminator sequencing chemistry is a labor intensive process requiring up to two hours of hands on time per day. Several robots have been evaluated for their ability to automate terminator chemistry.

The Qiagen 9600 BioRobot, Packard Multiprobe II and a custom in-house robot have been considered. The increase of throughput made possible with the 3700 DNA Analyser has increased the need for an improved data management system. A commercial automated data management and analysis system, PE Informatics' BioLIMS, is being evaluated. BioLIMS is a centralized relational database with an open modular architecture. BioLIMS can reduce the need for multiple copies of files, is searchable and can be set up to automate tasks such as data trimming, back-up and archiving. Sequencher for BioLIMS is the version of Gene Codes' DNA Analysis software package that allows integration with BioLIMS. Performance and features of the Sequencher software program will be presented.

P-182
Estimation of the Confidence Limits of Oligonucleotide Array-Based Measurements of Differential Expression

Paul K. Wolber, Andrew S. Atwell, Cynthia Y. Enderwick, Glenda C. Delenstarr, Andreas N. Dorsel, Karen W. Shannon, Robert H. Kincaid, Chao Chen, Shad R. Schidel & Michael P. Aschoff, Agilent Technologies, Palo Alto, CA

Microarrays of oligonucleotide probes can be used to simultaneously infer the differential expression states of many mRNA's in two samples. Such inferences are limited by systematic and random measurement errors. Systematic errors include signal gradients, imperfect feature morphologies, mismatched sample concentrations, cross-hybridization and scanner bias. Random errors arise from chemical and scanning noise, particularly for low signals. We have used a combination of two-color labeling (with fluor xchange) and rational array design to minimize systematic errors from gradients, imperfect features and mismatched sample concentrations. On-array specificity control probes and careful probe design were used to correct for cross-hybridization. Random errors were reduced via automated bad feature flagging and an advanced scanner design. We have scored feature significance, using established statistical tests. We have then estimated the intrinsic random measurement error as a function of average probe signal via sample self-comparison experiments (human K-562 cell mRNA). Finally, we have estimated the accuracy of differential expression measurements between K-562 cells and HeLa cells by evaluating the consistency with which different probes to the same mRNA measure differential expression. The data establish the importance of the use of sensitive probes and the elimination of systematic errors in producing reliable estimates of differential expression.

P-183
Negative Selection of Intact mRNA for Full-Length cDNA Library Construction

Ning Wu, Troy Moore, Shannon Wang, MaryAnn Taylor, Dewight Cowley II, Mandy B. Hammons, Barry S. Fitts, Leslie A. Crow, Monaz V. Baria, Jeneco S. Thomas, James R. Hudson, Jr, Research Genetics, Inc., Huntsville, AL

In the process of generating full-length cDNA library, the selection of intact mRNA is the essential step. Many methods have been developed focusing on the manipulation of the intact 5' end cap structure of the mRNA (i.e. "Oligo-capping", "Cap trapper", etc.). We have developed a novel method for intact mRNA selection based on the elimination of uncapped RNAs. A negative selection strategy that removes both uncapped mRNA and other non-mRNA molecules that present a phosphate at the 5' end has been applied in the mRNA purification procedures. A biotinylated oligoribonucleotide (r-oligo) is ligated to the 5' end phosphate by utilizing T4 RNA Ligase. By using streptavidin extraction and phenol/chloroform purification, all truncated mRNA and non-mRNA are removed from the intact mRNA. We have applied this methodology in construction of a mouse brain cDNA library. The sequence analysis of 137 clones revealed that 15% of clones displayed no matches in both "NR" and "dB EST" databases. 47% of the clones are known genes and within them, the full length clones total more than 68%. 5' ends of the known genes analyzed are from -305 to +196, further sequence data is under analysis.

P-184
Dynamically Organizing Gene Family Research Literature for *Arabidopsis thaliana*

Dongying Wu, Daniel Haft, Maria-Ines Benito, Owen White, The Institute for Genomic Research, Rockville, MD

Whole genome-scale gene family analysis is one of the most important approaches for understanding protein

STIC-ILL

05/23

447317

From: Zeman, Mary
Sent: Friday, May 23, 2003 11:26 AM
To: STIC-ILL
Subject: References 09/961058

Please send a copy of the following references
Thank you
Mary K. Zeman
Examiner, 1631
305-7133
CM1 12A17
mailbox: CM1 12D01

SO Methods in Molecular Biology (Totowa, New Jersey) (2000), 132(Bioinformatics Methods and Protocols), 71.91
CODEN: MMBIED; ISSN: 1064.3745

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 1.7 ISSN: 1415.

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 9.15 ISSN: 1415.4757

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 17.23 ISSN: 1415.4757

SO International Genome Sequencing and Analysis Conference, (2000) Vol. 12, pp. 97.98. print.
Meeting Info.: 12th International Genome Sequencing and Analysis Conference Miami Beach, Florida, USA September 12.15, 2000

Genetics and Molecular Biology, 24 (1-4), 9-15 (2001)

Bioinformatics of the sugarcane EST project

Guilherme P. Telles, Marília D.V. Braga, Zanoni Dias, Tzy-Li Lin, José A.A. Quitzau, Felipe R. da Silva and João Meidanis*

Abstract

The Sugarcane EST project (SUCEST) produced 291,904 expressed sequence tags (ESTs) in a consortium that involved 74 sequencing and data mining laboratories. We created a web site for this project that served as a 'meeting point' for receiving, processing, analyzing, and providing services to help explore the sequence data. In this paper we describe the information pathway that we implemented to support this project and a brief explanation of the clustering procedure, which resulted in 43,141 clusters.

INTRODUCTION

The application of expressed sequence tag (EST) technology has proven to be an effective tool for gene discovery (Adams *et al.*, 1991), gene mapping (Schuler, 1997) and the generation of gene expression profiles (Boguski and Schuler, 1995).

EST projects are usually conducted by a single laboratory, which prepares the cDNA libraries, isolates and sequences clones, analyzes the data and submits it to GenBank. However, the Sugarcane EST project (SUCEST) involved the cooperation of 24 sequencing laboratories, a bioinformatics laboratory, a coordinating laboratory, 50 data mining groups scattered throughout Brazil and an international relations group. A new Brazilian bioinformatics group also became associated with the project during a later phase. Starting early in 1999, in 15 months the SUCEST project generated 291,904 sequences from 260,352 clones from 37 different libraries.

Brazilian genome research has been consortium-based since its first project, the sequencing of the complete genome of the phytopathogenic bacterium *Xylella fastidiosa* (Simpson *et al.*, 2000), conducted by the Organization for Nucleotide Sequencing and Analysis (ONSA network). Although a consortium-based genome project provides a larger number of researchers, technicians and sequencing machines it demands a much more organized data flow. In the SUCEST project, the Bioinformatics Laboratory (Laboratório de Bioinformática - LBI) was responsible for receiving data from a network of sequencing laboratories, assessing quality, storing and clustering the data, and providing many other services. In this paper these tasks are described in some detail and quantitative figures from the project are given.

METHODS

Computational systems

For a short time in the beginning of the project, the SUCEST web site was hosted by a personal computer with 128 MB of memory running the Linux operating system (Red Hat 6.2) but now the site resides on a Compaq AlphaServer ES40 with two Alpha 667 MHz processors, 8 GB of RAM and 384 GB of hard-disk storage space running OSF-1 operating system version 4.0G. However, the bulk of the project was executed on a Compaq AlphaServer DS20 with two Alpha 500 MHz processors, 4 GB of RAM and 144 GB of hard-disk storage space running OSF-1 version 4.0F. Since this was the system on which most of the tools were developed we will concentrate on it for the rest of the paper.

The Web engine server is Apache (www.apache.org) version 1.3.9. Programs were written in Perl version 5.005 (www.cpan.org), and PHP version 3.0.12 (www.php.net). The database management system is MySQL version 3.22.26a (www.mysql.com).

Input data consisted of data received through web forms, including chromatograms produced by ABI 377 sequencing machines (Applied Biosystems), and data mining reports in HTML format.

The base calling and sequence extraction programs used were phred version 0.980904.e (www.phrap.org) and phd2fasta version 0.990622.d (www.phrap.org). The sequence comparison programs used were cross-match version 0.990319 (www.phrap.org) and blastall version 09/19/1999 (www.ncbi.nlm.nih.gov) that implements the BLAST algorithm (Altschul *et al.*, 1997). Assembly programs were phrap version 0.990319 (www.phrap.org) and CAP3 (Huang and Madan, 1999). Off-the-shelf scripts

Bioinformatics Laboratory - IC-UNICAMP, C.P. 6176, 13083-970 Campinas, SP, Brazil.
Send correspondence to Guilherme P. Telles. E-mail: pimentel@ic.unicamp.br.

were used to provide search by keywords in the reports produced by data mining groups, database administration and other minor tasks. Each piece of software used is either free for academic purposes or was developed by our team.

Computational methods

From a computational point of view, SUCEST may be seen as a large data repository and as a provider of Internet-based services for a community of different users. Figure 1 shows the major relationships between users, services, data and programs in the project.

There are several types of users: members of sequencing laboratories who submit chromatograms from clone libraries, members of data mining laboratories who perform searches on the project database and publicize their results in data mining reports, and members of the project coordination team who monitor the status of the project and the distribution and validation of control plates. These users interact with data through services that add to, retrieve from, and update the data repositories.

Data include sugarcane ESTs, information about project members, data mining reports, control data, summaries and the output from programs that perform automated searches in databases, organize the sequences into clusters and the clusters into categories. In the following paragraphs we describe the users, data, and SUCEST services and programs, showing how they interact.

DEFINITIONS

Objects

In the SUCEST project data is stored in two different kinds of repositories: operating system directories and a relational database. The directories hold biological sequence files, results from BLAST and cross-match searches in biological databases, and data mining reports. Biological sequence files include chromatograms, files in a standard format called fasta format (www.ncbi.nlm.nih.gov/BLAST/fasta.html), quality files, and files generated by clustering, categorization and comparative genomics procedures. The project uses only one relational database, with several interconnected tables that store other biological and management data, *e.g.* libraries, sequencing plates and data on laboratories and their members. The database also points to data in directories. The major entities (objects) in our database are described below, where we also introduce quantitative figures and details from the project's pipeline.

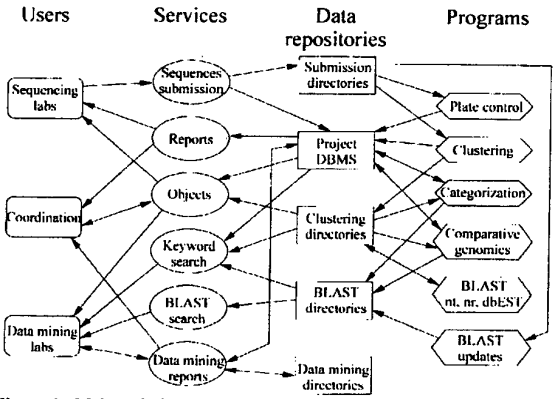

Figure 1 - Major relationships between users, services, data and programs involved in the SUCEST project. Arrows indicate the flow of information.

Laboratories

There are 78 laboratories involved in the SUCEST project that belong to one or more of five groups: the DNA Coordination Group, the Bioinformatics Group, the Data Mining Group, the Sequencing Group and the International Cooperation Group. Each participating laboratory is identified by a two-letter code. The services and data that a member of a particular laboratory can access depend on the group to which the laboratory belongs. A member of each laboratory is designated as being the head of the unit involved in SUCEST-related work and receives notification of some of the activities performed by the laboratory members.

Members

A SUCEST member is a person who belongs to at least one laboratory. Several members belong to both a sequencing laboratory and a data mining laboratory. Data held on members include their name, the laboratories to which they belong, their e-mail address, phone numbers and a login name and password to grant access to authorized services. SUCEST had 256 members as at March 25, 2001.

Libraries

The ESTs included in the SUCEST database came from 37 different libraries prepared from different sugarcane tissues under different conditions (Vettore *et al.*, 2001). The name and description of the library and vector employed in cloning were recorded for each library. Each library received a two-letter code indicating the tissue from which the library was derived, together with a consecutive number assigned for every new library derived from the same tissue. For example, LR1 indicates that the library came from leaf roll (LR) with long inserts (library 1) while LR2 shows that the library came from leaf roll (LR) with small inserts (library 2). There are three possibilities for the status of each library: 'test' for validating libraries, 'start' for libraries released for sequencing and 'stop' when the DNA Coordination Group decides it is not worth continuing to sequencing a distributed library. Of the 37 libraries prepared for the project, 32 were started and 5 were aban-

doned after the 'test' phase. Those not formally started either produced too much redundancy or very small reads.

Plates

SUCEST clones are organized in 96-well plates that hold clones from the same library in an 8 x 12 grid. Sequencing is done for a whole plate and the data is sent to the LBI for processing and storage. Data for a plate include the library that it came from and the laboratory that is authorized to send data on this plate. A plate has a three-digit identification tag, except for control plates (see below), which have the letter 'C' and two digits. The SUCEST database holds data from 2,771 different plates.

Reads

Reads are the same as ESTs and are extracted using the phred program from chromatograms submitted by the sequencing laboratories and screened for vectors with the cross-match program. All reads are stored in directories as chromatogram files and also as a pair of text files holding the sequence and its quality in fasta format. For every read the following attributes are stored in the database: the plate and the position on the plate where the read came from; information about the submission process (*e.g.* date and time of submission); the number of vector and non-vector bases with phred quality equal to or higher than 20; the number of vector and non-vector bases with phred quality less than 20; the starting and ending positions for every vector sequence identified in the read and whether or not the read has relevant data (see preparation sheet below.)

Every read has a name that is a concatenation of its laboratory, library and plate codes, plate position and read direction (5' or 3'). For example, reading from right to left, the string SCACAD1001A01.g is the name for the 5' read (3' uses .b as a suffix.) of the clone in well A01 of plate 001 of library AD1, sequenced by laboratory AC. The prefix SC stands for sugarcane. Every position on the plate is identified by its row (A to H) and column (01 to 12).

Preparation sheet

Before a laboratory can sequence and submit a plate, it must provide a sheet of information about the process used to prepare the plate. There are records in the database for every well where bacteria did not grow and for the wells from which it was not possible to obtain DNA. Every well marked as a problem corresponds to a sequence without information relevant to the project.

Control plates

For every set of 12 plates a control plate is built using the 8th column of each controlled plate, so 12 columns make one control plate that is sequenced. The sequences from both control and controlled plates are compared against each other using cross-match, and the matches are stored in the database. A criterion, based on the matches distribution over the control and controlled plates, was established to automatically mark plates that probably had tracking and naming errors due to plate preparation and sequencing processes. Matches distributions could be visualized via a web service, and plates with problems could be fixed and resubmitted by the laboratory that produced them.

Clusters

SUCEST reads are grouped by the clustering procedure described below, which creates sets of aligned reads that we call clusters. In our database we store the reads that are part of each cluster. Moreover, in addition to being a set of reads, a cluster has an alignment and a consensus sequence. Alignments, consensus sequences, and quality files are stored in cluster directories. A cluster also has a name, which is equal to the name of oldest read in the cluster.

Services and programs

Data enter and are retrieved from the SUCEST data repository through a set of services available on web pages hosted at LBI. Data is also generated within the LBI by programs that are executed either automatically or manually. Brief descriptions of these services and programs are presented below and provide a general overview on how the SUCEST web site is organized and how it works.

Data retrieval

Data is retrieved from the SUCEST database in units called 'objects' which are the same as the data entities described above under 'Definitions'. Each object has its own web page containing information about the object and links to any other object, service or report directly related to it. Starting from a laboratory or library object it is possible to reach the web page of any other object. Some objects point to pages that include data extracted from the directory structure of the project. For instance, one can visualize reads and its qualities in many versions: immediately after submission but before screening, after screening but before trimming (see below under 'Clustering and Trimming') and after trimming. For clusters, it is possible to see the reads in a cluster and their alignments, including the consensus.

An *object search* service was created to allow direct access to any object. Given the code and the type of the object, the service delivers its page. For the 'Member' object type it is possible to search by name, email, department, city or institution.

Besides objects, some reports that summarize data are also available for the project: the *Summary of Submitted Reads* gives totals per laboratory or per library of submitted, payable and clusterizable reads, and *the Summary of Control Plates* gives the totals of accepted and rejected control plates.

SUCEST database users who are SQL (Structured Query Language) literate may take advantage of a service that allows generic queries to the database. Queries can be typed in a web form and the results are returned in tabular fashion. Entity-relationship diagrams and table descriptions for our database are available to help users in this task.

Sequences submission

Sequences are submitted by sequencing laboratories only, the submission process requiring the user to access the project's web site using a valid login/password pair to upload a set of 96 chromatograms (*i.e.* one plate). When an upload finishes certain pre-requisites are verified: all chromatograms must belong to the same plate, the laboratory that is trying to submit a plate must be the one authorized to do so, the preparation sheet for that plate must have already been submitted and the reads must be in accordance with the naming conventions.

If the pre-requisites are satisfied, the phred and phd2fasta programs are used to extract the sequences and their qualities in fasta format from chromatograms and the cross-match program is used to mask vector sequences in the reads. These steps take only a few minutes (this time has essentially been constant during the project because the analysis done upon submission does not depend on the other reads present in the repositories).

After submission analysis, a report that summarizes the process and the sequences received is presented to the submitter who is asked to confirm the submission or not. If the submission is confirmed, the database is updated and if there is an older version of the plate it is replaced. Directories are updated as well. If the submission is not confirmed (*e.g.*, if the submitter is not happy with the quality assessment) the submission is discarded.

Figure 2 shows the path followed by a read in the LBI, starting from the submission. The submission procedure corresponds to the part of the figure starting at 'Zip file', extending through top line and reaching the 'Report Generator'. Other steps in the diagram are performed by programs described in the following sections.

Clustering and trimming

Clustering of ESTs is important to reduce the amount of sequence data that miners have to look at, and to orga-

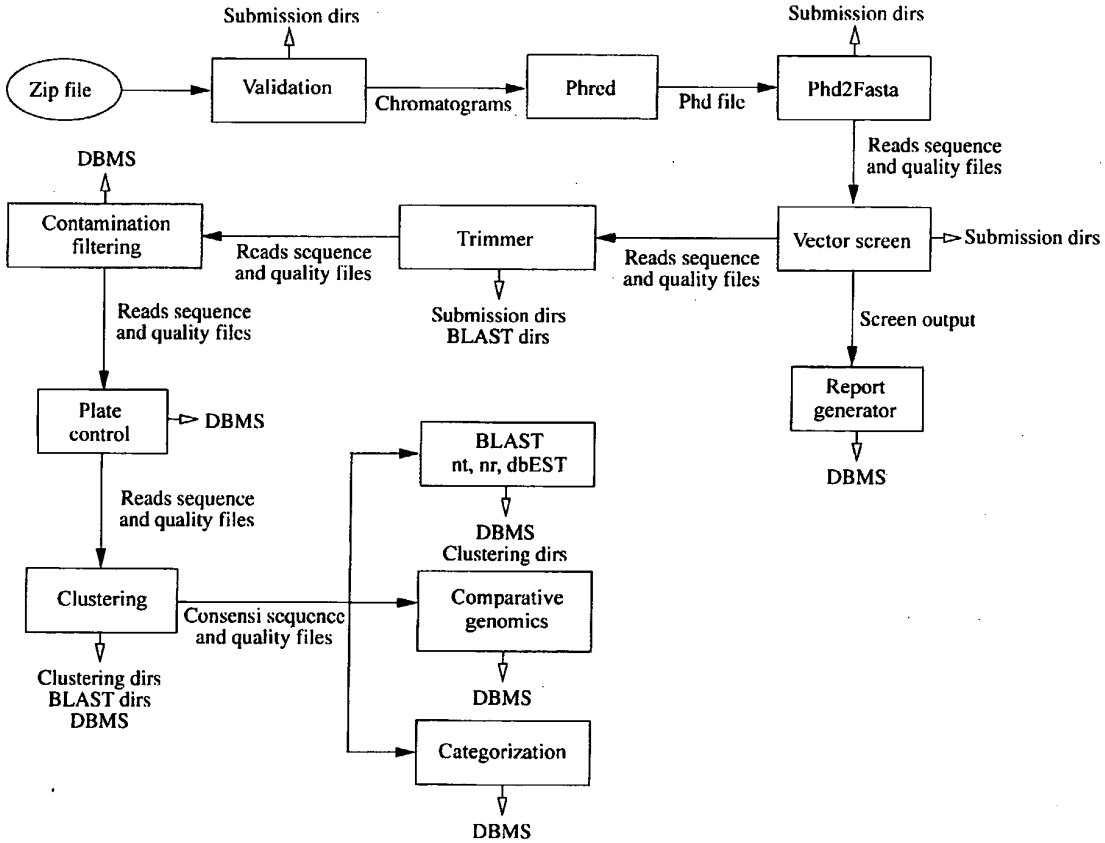

Figure 2 - The operations taking place on a read in the SUCEST pipeline. Black arrows linking boxes indicate data that flow from one stage to the next while white-headed arrows going out of boxes indicate date repository updates.

nize the reads in a less redundant set. In the SUCEST database, clustering had as an additional motive the need to estimate the level of redundancy in the libraries.

Early on two pivotal decisions were made, the first being that each cluster should reflect a transcript rather than a gene, allele or other biological entity while the second was that a cluster consists not only of a set of reads but also of an alignment of these reads.

In this context, our first scheme was to group similar transcripts and to produce consensus sequences using the assembly program phrap. This strategy was sufficient in the early stages of the project but, as data accumulated, a series of problems forced us to change the scheme, as described below.

To minimize artifacts, reads were trimmed before clustering. This trimming procedure started with vector masking using the cross-match program followed by removal of some of the poly-A, vector and adapter regions. A quality trimmer was also applied, removing bases from the ends of the sequence one by one until there were at least 12 bases with phred quality above 15 in a window of 20 bases at the end. Reads were also checked for contamination against *Xylella fastidiosa*, *Xanthomonas citri*, *Escherichia coli* and other potential contaminants that could be present in the laboratories that produced the libraries. BLAST was used to compare the reads and potential contaminators and if a match of at least 100 bases and more than 90% identity occurred the read was marked as probably being due to contamination. However, marked reads were kept in clustering and subsequent analyses to allow data miners to decide for themselves whether or not a specific read was contaminated.

Trimmed reads were assembled using the phrap program with quality and stringent arguments (-penalty -15 -bandwidth 14 -minscore 100 -shatter_greedy). Every contig and singlet produced by phrap was taken as a cluster. As new plates came in, a program automatically updated the database, directories and BLAST results for every cluster that changed and was already in the database. Initially, clustering was performed every day but as the set of sequences grew the updates became sparser, running once a week. In the final phases of the project, clustering would typically occupy an entire processor for about 20 hours.

The last assembly done with phrap included 261,609 trimmed reads and produced 81,223 clusters. However, changes were made due to remarks made by several members of the project that the total number of clusters in the database was unreasonably large, that many clusters were malformed and that some clusters appeared as if they could be combined. These changes are described in detail by Telles and da Silva (2001). The new scheme was based on careful testing and evaluation, and consisted of a more elaborate trimming procedure, the use of the CAP3 assembler (Huang and Madan, 1999), which is the same tool used to produce TIGR's gene indices (Quackenbush *et al.*, 2000). Trimming in this new procedure included ribosomal RNA removal, comprehensive removal of poly-A, poly-T, vector and adapter regions and improved low-quality-end trimming. CAP3 was fed with 237,954 reads and their quality data and produced 43,141 clusters.

Both clustering versions are accessible through the project web site, with data from both methods available for most services.

Keyword search

Keyword search is a service that allows users to search for a set of keywords in the header lines of every sequence in NCBI's nr, nt and dbEST databases (www.ncbi.nlm.nih.gov) that hits any cluster in SUCEST. To perform a query the user gives a database name (nr, nt or dbEST), a logical expression of keywords (that may include 'or' and 'and' connectors) and the maximum e-value required (an optional parameter which defaults to 1e-5 = 10^{-5}). The service then returns the clusters that have a hit with the expected or better e-value, and whose subject heading contains words satisfying the logical expression. The resulting list of clusters is ordered by e-value.

A program was created for keeping BLAST results against nr, nt and dbEST up to date for all SUCEST clusters. A BLAST result against a certain database is considered *outdated* for a SUCEST cluster if the cluster was newer than the result or if the cluster or the database were modified after the last BLAST run. When the program finds outdated BLAST results it builds a queue giving priority to older clusters. If the databases are on different computers the system is able to reduce the processing time by running several BLASTs in parallel (one on each remote server) and takes about 2 or 3 days. If the databases are on a single computer, BLAST searches take considerably longer.

Subclustering

This service is used to evaluate statistics about subsets of clusters obtained by clustering, including read frequency by cluster size, total reads, total clusters, redundancy and novelty. To select the subset of clusters, the user has to indicate the reads that belong to the clusters. Any cluster that contains a read in the selection is included in the evaluation. To locate reads, one or more elements (laboratory, library, plate, position and direction) from their names should be selected, *e. g.* selecting a particular laboratory will generate the statistics for the clusters that have at least one read sequenced by that laboratory.

BLAST search

A BLAST service allows searches against SUCEST reads, reads in their trimmed version and cluster consensi. These databases were updated automatically on a daily basis to incorporate new reads and consensi.

Data mining report

Data mining groups submit HTML formatted reports to the SUCEST site and update them periodically. Users may access reports through an index page that provides access to the reports of every data mining group and a keyword search is also available. When a report archive is uploaded a service takes care of unpacking the files and updating the index page and the search index. Information about reports is also kept in the SUCEST database, including the name and a summary of the project, its members and a submission date and submitter name.

Categorization

SUCEST members tried to categorize the clusters in the project, in an attempt to determine their function and to aggregate information. Thirty categories were defined, and 32,438 proteins with known function were selected from public databases to serve as examples in each category. Public databases included MIPS *Arabidopsis thaliana* database (mips.gsf.de), Clusters of Orthologous Groups - functional annotation (www.ncbi.nlm.nih.gov/COG/), EGAD cellular roles (www.tigr.org/docs/tigr-scripts/egad_scripts /role_report.spl) and others.

Categorization was achieved by two methods: automatic and manual. In automatic categorization a database was constructed containing the proteins selected from public databases and a BLAST search was performed against this database using SUCEST clusters as input. Any cluster was considered to be in category X if it matched a category X sample protein with an e-value better than or equal to 10^{-10} and covered 70% or more of the example. A cluster could be in many different categories. This method categorized 36% of the 43,141 clusters. For manual categorization a web service was built to allow manual annotation when automatic annotation produced ambiguous categorization or produced no categorization at all. Based on BLAST results against the nr database, SUCEST members were able to establish a direct relation between a cluster and a category. Manual annotation significantly increased the number of categorized clusters and as of March 20th, 2001, 60.5% of the clusters were categorized.

Comparative genomics

To obtain information on sugarcane and its relationship to other species, SUCEST cluster consensi were compared against other organisms. The first organism selected for comparison was the model plant *Arabidopsis thaliana*. Every cluster consensus was BLASTed against *A. thaliana* chromosomes, proteins and ESTs. Clusters that produced no matches against *A. thaliana*, were also BLASTed against ESTs from *Lycopersicon esculentum*, *Glycine max*, *Lotus japonicus*, *Hordeum vulgare*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Triticum aestivum* and *Medicago truncatula*. Results from these searches were inserted in our database, allowing queries to determine the distribution of these hits per library, per cluster, or some other grouping criteria.

Management

These services provide a way for the DNA Coordination Group to input management information into the SUCEST database. This information is used mainly by services that perform checking and summarizing operations. Using the library management services, the DNA Coordination Group modifies the status of any library and assigns plates to sequencing laboratories. Manual plate approval is also possible via a service that displays control and controlled plates showing which cells match in control and controlled plates.

DISCUSSION

A key aspect of the project was the close interaction between the biological laboratories and the LBI. Discussion lists or telephone calls were used so that users could give suggestions for new services and quickly point out problems with the services (broken links, bugs, etc.) This daily, intensive interaction was undoubtedly one of the main reasons for the success of the project.

Clustering started early and had a dramatic impact during the project. Re-clustering on a regular basis demanded designing and implementing programs to update databases and BLAST results against the nr, nt and dbEST databases, and also used a lot of processor time. When another clustering scheme was adopted the web site had to change to accommodate both versions simultaneously and to show relationships between clusters in different versions and both bioinformatics and data mining staff needed some time to adapt to the changes.

The two most important lessons learnt during the SUCEST project were 'avoid changing systems' and 'keep reference sequences, not cluster lists' which we will discuss in more detail in the following paragraphs.

Avoiding changes in the systems is important. During this project we had to change the underlying computing system twice, the first time from a personal computer to a medium-sized server and then from this to a larger server. These changes caused many problems, *e.g.* programs that used to work on one system would not work on the other system, users had to get used to new addresses etc. The migration process proved time-consuming and error-prone. Our advice would be to set up a system that is big enough right from the start and keep the project there for as long as possible. To minimize the impact of migration it is important to devise the directory structure in a system-independent way, for instance data can be placed in directories that will not conflict with system directories and programs can be installed in standard locations and execution path variables used to assure they will work. Another important piece of advice is to use software that combines many phys-

ical disks into one big volume of, say, a few hundred gigabytes. Most vendors provide such software for a small fee.

It is also important to keep reference sequences instead of lists of clusters. In this project, data accumulated at a fast rate and clustering was redone frequently. Some data mining groups had problems trying to keep up with the frequent updates because they maintained lists of relevant clusters. Each time the clustering was redone some clusters would disappear (merge into larger ones) or the read composition of a cluster would change, requiring a lot of manual labor. Our advice would be to use reference sequences from Genbank or another stable sequence database, which can then be used as queries to retrieve the cluster lists via BLAST. Proceeding in this way lists can be quickly reconstructed from the reference sequences using automated methods.

There are many other programs, not presented here, that contribute to the functionality of the SUCEST web site. Some services and programs have already been disabled (*e.g.* the sequence submission and plate control programs) but others, such as the keyword search, BLAST and report submission programs are still being used by data mining laboratories and will be used by the international community when the web site goes public. This will certainly transform the meeting point of the project's community into the meeting point of a wider group which will produce new demands for services and data storage.

ACKNOWLEDGMENTS

This work was supported by FAPESP, CNPq and COPERSUCAR.

RESUMO

O projeto SUCEST (Sugarcane EST Project) produziu 291.904 ESTs de cana-de-açúcar. Nesse projeto, o Laboratório de Bioinformática criou o *web site* que foi o "ponto de encontro" dos 74 laboratórios de sequenciamento e *data mining* que fizeram parte do consórcio para o projeto. O Laboratório de Bioinformática (LBI) recebeu, processou, analisou e disponibilizou ferramentas para a exploração dos dados. Neste artigo os dados, serviços e programas implementados pelo LBI para o projeto são descritos, incluindo o procedimento de *clustering* que gerou 43.141 *clusters*.

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. (1991). Complementary DNA sequencing: "expressed sequence tags" and the human genome project. *Science 252*: 1651-1656.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*: 3389-3402.

Boguski, M. and Schuler, G. (1995). ESTablishing a human transcript map. *Nature Genetics. 10*: 369-371.

Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res. 9*: 868-877.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res. 28* (1): 141-145.

Schuler, G. (1997). Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J Mol Med. 75* (10): 694-698.

Simpson, A.J.G., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R., Alves, L.M.C., Araya, J.E., Baia, G.S., Baptista, C.S., Barros, M.H., Bonaccorsi, E.D., Bordin, S., Bove, J.M., Briones, M.R.S., Bueno, M.R.P., Camargo, A.A., Camargo, L.E.A., Carraro, D.M., Carrer, H., Colauto, N.B., Colombo, C., Costa, F.F., Costa, M.C.R., Costa-Neto, C.M., Coutinho, L.L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A.P., Ferreira, A.J.S., Ferreira, V.C.A., Ferro, J.A., Fraga, J.S., Franca, S.C., Franco, M.C., Frohme, M., Furlan, L.R., Garnier, M., Goldman, G.H., Goldman, M.H.S., Gomes, S.L., Gruber, A., Ho, P.L., Hoheisel, J.D., Junqueira, M.L., Kemper, E.L., Kitajima, J.P., Krieger, J.E., Kuramae, E.E., Laigret, F., Lambais, M.R., Leite, L.C.C., Lemos, E.G.M., Lemos, M.V.F., Lopes, S.A., Lopes, C.R., Machado, J.A., Machado, M.A., Madeira, A.M.B.N., Madeira, H.M.F., Marino, C.L., Marques, M.V., Martins, E.A.L., Martins, E.M.F., Matsukuma, A.Y., Menck, C.F.M., Miracca, E.C., Miyaki, C.Y., Monteiro-Vitorello, C.B., Moon, D.H., Nagai, M.A., Nascimento, A.L.T.O., Netto, L.E.S., Nhani Jr., A., Nobrega, F.G., Nunes, L.R., Oliveira, M.A., de Oliveira, M.C., de Oliveira, R.C., Palmieri, D.A., Paris, A., Peixoto, B.R., Pereira, G.A.G., Pereira Jr., H.A., Pesquero, J.B., Quaggio, R.B., Roberto, P.G., Rodrigues, V., de M. Rosa, A.J., de Rosa Jr., V.E., de Sa, R.G., Santelli, R.V., Sawasaki, H.E., da Silva, A.C.R., da Silva, F.R., da Silva, A.M., Silva Jr., W.A., da Silveira, J.F., Silvestri, M.L.Z., Siqueira, W.J., de Souza, A.A., de Souza, A.P., Terenzi, M.F., Truffi, D., Tsai, S.M., Tsuhako, M.H., Vallada, H., Van Sluys, M.A., Verjovski-Almeida, S., Vettore, A.L., Zago, M.A., Zatz, M., Meidanis, J. and Setubal, J.C. (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature 406*: 151-157.

Telles, G.P. and da Silva, F.R. (2001). Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology 24* (1-4): 17-23.

Vettore, A., da Silva, F.R., Kemper, E. and Arruda, P. (2001). The libraries that made SUCEST. *Genetics and Molecular Biology 24* (1-4): 1-7.

STIC-ILL

105/23

447318

From: Zeman, Mary
Sent: Friday, May 23, 2003 11:26 AM
To: STIC-ILL
Subject: References 09/961058

Please send a copy of the following references
Thank you
Mary K. Zeman
Examiner, 1631
305-7133
CM1 12A17
mailbox: CM1 12D01

COMPLETED

SO Methods in Molecular Biology (Totowa, New Jersey) (2000), 132(Bioinformatics Methods and Protocols), 71.91
CODEN: MMBIED; ISSN: 1064.3745

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 1.7 ISSN: 1415.

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 9.15 ISSN: 1415.4757

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 17.23 ISSN: 1415.4757

SO International Genome Sequencing and Analysis Conference, (2000) Vol. 12, pp. 97.98. print.
Meeting Info.: 12th International Genome Sequencing and Analysis Conference Miami Beach, Florida, USA September 12.15, 2000

Genetics and Molecular Biology, 24 (1-4), 1-7 (2001)

The libraries that made SUCEST

André L. Vettore[1,a], Felipe R. da Silva[1,b], Edson L. Kemper[1,c] and Paulo Arruda[1,4]

Abstract

A large-scale sequencing of sugarcane expressed sequence tags (ESTs) was carried out as a first step in depicting the genome of this important tropical crop. Twenty-six unidirectional cDNA libraries were constructed from a variety of tissues sampled from thirteen different sugarcane cultivars. A total of 291,689 cDNA clones were sequenced in their 5' and 3'end regions. After trimming low-quality sequences and removing vector and ribosomal RNA sequences, 237,954 ESTs potentially derived from protein-encoding messenger RNA (mRNA) remained. The average insert size in all libraries was estimated to be 1,250bp with the insert length varying from 500 to 5,000 bp. Clustering the 237,954 sugarcane ESTs resulted in 43,141clusters, from which 38% had no matches with existing sequences in the public databases. Around 53% of the clusters were formed by ESTs expressed in at least two libraries while 47% of the clusters are formed by ESTs expressed in only one library. A global analysis of the ESTs indicated that around 33% contain cDNA clones with full-length insert.

INTRODUCTION

Single-pass sequencing of cDNAs to generate "expressed sequence tags" (ESTs) has proven to be a powerful, economical and rapid approach to identify genes that are preferentially expressed in certain tissue or cell types of multicellular organisms (Adams *et al.*, 1991, Hwang *et al.*, 1997, Liew *et al.*, 1994, Adams *et al.*, 1995). Increasing importance has also been attributed to ESTs as a tool for the annotation of complete genome sequences of mammalians and plants. Unique ESTs provided biological evidence of hundreds of predicted genes, newly discovered genes, or transcript isoforms leading to considerable advance in gene identification mission in multicellular organisms (Andrews *et al.*, 2000). Today, more than ten million ESTs are currently available through the dbEST entry of GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_sumary.html); however, only 14% of dbEST release 022301 of February 23, 2001 corresponds to plant sequences.

Another useful aspect of ESTs is in accessing genetic information of species with a complex genome, whose access is difficult using conventional genetics. This is the case of sugarcane, an important crop that is cultivated in the tropics for its high sucrose accumulation in the stalk. Among the cultivated crops, sugarcane possesses perhaps one of the most complex genomes (for a review see Grivet and Arruda, 2002). Modern sugarcane cultivars are hybrids derived from the crossing of *Saccharum officinarum*, usually having 2n = 80 chromosomes and *Saccharum spontaneum*, 2n = 40 - 128 chromosomes. In view of the structural differences between chromosomes of the two species, the hybrids possess different proportions of chromosomes, varying chromosome sets and complex recombinational events (Grivet and Arruda, 2002). This imposes tremendous difficulties in applying conventional plant breeding techniques to sugarcane.

As a first step in depicting the sugarcane genome, the ONSA consortium (Simpson and Perez 1998) launched in September of 1998 the Sugarcane Expressed Sequence Tag project (SUCEST), aiming at sequencing random ESTs and identifying around 50,000 unique genes (http://sucest.lad.ic.unicamp.br/en/).

To improve the probability of getting a maximum number of different ESTs, researchers have been using normalized and/or subtracted cDNA libraries that bring the frequency of each clone in a cDNA library within a narrow range (Soares and Bonaldo 2000). However, normalization and/or subtraction procedures are in general laborious and have the tendency of increasing the proportion of small insert clones. In the SUCEST project we have implemented an efficient procedure to generate conventional cDNA libraries to generate large scale ESTs from sugarcane. This paper describes the construction of these libraries, representing all major organs, harvested at different developmental stages and used to generate one of the largest plant EST collections.

[1]*Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil and Depto de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil.*

Present address:

[a]*Instituto Ludwig de Pesquisa sobre o Câncer, 01509-010, São Paulo, SP, Brazil.*

[b]*Embrapa Agrobiologia, BR 465, Km 47, CxP 74505, 23851-970, Seropédica, RJ, Brazil.*

[c]*Monsanto Company, GG5B, 700 Chesterfield Pkwy N, 63038, Chesterfield - MO, USA.*

Send correspondence to Paulo Arruda. E-mail: parruda@unicamp.br.

MATERIAL AND METHODS

Plant material

Sugarcane tissues were obtained from commercial cultivars (Table I) grown at the Copersucar experimental station (Piracicaba, SP, Brazil), at the Universidade Federal de São Carlos experimental station (Serra do Ouro, AL, Brazil) and at the Centro de Biologia Molecular e Engenharia Genética (Campinas, SP, Brazil). After harvesting, tissues were frozen in liquid nitrogen and stored at -80 °C.

RNA isolation

Total RNA was isolated using Trizol (Invitrogen) according to manufacturer's instructions. Due to the high carbohydrate content and the presence of phenolic compounds, total RNA from immature seeds was isolated according to the method described by Manning (1991).

Poly(A)$^+$ mRNA was purified from total RNA using Oligotex-dT (Qiagen) according to manufacturer's instructions. Purity and RNA integrity were assessed by absorbance at 260/280 nm and agarose gel electrophoresis.

cDNA library construction

Libraries were constructed using the SuperScript cDNA Synthesis and Plasmid Cloning Kit (Invitrogen) according to the manufacturer's protocols. One microgram of poly(A) + mRNA was reverse-transcribed using a poly-dT primer containing the *Not*I site. The efficiency of cDNA synthesis was monitored with radioactive nucleotides. The second cDNA strand was then synthesized by replacing the RNA in the hybrids with DNA by using a combination of RNase H, DNA Polymerase I and DNA Ligase. After the second-strand synthesis and ligation of *Sal*I adapters, cDNA was digested with *Not*I, generating cDNA flanked by *Sal*I sites at 5' ends and *Not*I sites at the 3' ends. Excess adapters were removed and cDNAs were size fractioned in a 40 cm long 1 mm ID Sepharose CL-2B columm. One hundred and fifty µL fractions were collected and 8 µL aliquots of each fraction was electrophoresed in 1.5%

Table I - Description of the SUCEST Libraries.

Library code	Library name	Description	Sugarcane variety
AD1	*G. diazotroficans1*	Mixture of tissues from root to shoot zone, stem and apical meristem of plantlets cultivated *in vitro* and infected with *Gluconacetobacter diazotroficans*	P70-1143[3]
AM1, AM2	Apical Meristem	Apical meristem of young plants	SP80-3280[2]
CL6	Calli	Pool of *calli* treated for 12 h at 4 °C and 37 °C in the dark or ligth	SP80-3280[1]
FL1, FL3, FL4, FL5, FL8	Flower 1, 3, 4, 5 and 8	Flowers harvested at different developmental stages	SP80-87432[1] PB5211 x P57150-4[1]
HR1	*H. rubrisubalbicans1*	Mixture of tissues from root to shoot zone, stem and apical meristem of plantlets cultivated *in vitro* and infected with *Herbaspirilum diazotroficans*	SP70-1143[3]
LB1, LB2	Lateral Bud 1 and 2	Lateral buds from mature plants	SP80-3280[1]
LR1, LR2	Leaf Roll 1 and 2	Leaf roll from immature plants	SP80-3280[1]
LV1	Leaf 1	Etiolated leaves from plantlets grown *in vitro*	SP83-5077 SP80-185 SP87-396 SP80-3280 SP803280 x SP81-5441[1]
RT1, RT2, RT3	Root 1, 2 and 3	0.3 cm-length roots from mature plants and root apex	SP80-3280[1]
RZ1, RZ2	Root to shoot	Root to shoot zone of young plants	SP80-3280[1]
RZ3	zone 1, 2 and 3		
SB1	Stalk Bark 1	Stalk bark from mature sugarcane plants	SP80-3280[2]
SD1, SD2	Seeds 1 and 2	Developing seeds	CB47-89 RB855205 RB845298 RB805028[4]
ST1, ST3	Stem 1 and 3	First and fourth internodes of immature plants	SP80-3280[1]

cDNA libraries were constructed from different tissues sampled from different varieties grown at Copersucar experimental station (Piracicaba-SP)[1], CBMEG - Universidade Estadual de Campinas (Campinas-SP)[2], Universidade Federal do Rio de Janciro (Rio de Janeiro-RJ)[3], and Universidade Federal de São Carlos experimental station (Serra do Ouro-AL)[4].

agarose gel to determine the size range of cDNAs. Fractions with cDNAs with a minimum size of 500 base pair (bp) were pooled and ligated to pSPORT1 vector (Invitrogen) predigested with *SalI* and *NotI*. The resulting plasmids were transformed in DH10B cells (Invitrogen) by electroporation. Unamplified libraries were plated and individual colonies picked and transferred to 96 well plates containing liquid Circle Grow (CG) medium (BIO 101), supplemented with 100 mg/L of ampicillin and 8% glycerol. Three copies of each cDNA clone were stored at -80 °C.

Template preparation and DNA sequencing

DNA template preparations and sequencing reactions were performed in a 96-well format. Plasmid templates were prepared using modified alkaline lysis (http://sucest.lad.ic.unicamp.br). Sequencing reactions were performed on plasmid templates using a quarter of the standard volume of ABI Prism BigDye Terminator Sequencing Kit (Applied Biosystems) and the T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') that hybridizes upstream of the *SalI* site in the pSPORT1 polylinker (5'end of the cDNA inserts) or the SP6 promoter primer (5'-ATTTAGGTGACACTATAG-3') that hybridizes downstream of the *NotI* site (3'end of the cDNA inserts). Reaction products were precipitated with 95% ethanol using sodium acetate (3M) and Glycogen (1g/L) as carriers and washed twice with 75% ethanol before drying under vacuum. The sequencing reaction products were analyzed on 377-96 ABI Sequencers.

Sequence analysis

Sequencing of sugarcane ESTs was performed by 23 laboratories located in Universities and Research Institutes of the State of São Paulo and sequences were processed by the Bioinformatics laboratory (LBI) located at Instituto de Computação, Universidade Estadual de Campinas. A detailed description of the methods used to receive, process, analyze, and display the sequences along with additional tools to help explore the sequence data can be found in this issue (Telles *et al.*, 2001, Telles and da Silva, 2001).

RESULTS AND DISCUSSION

The SUCEST strategy

EST programs to acquire information about the transcriptome has been carried out for hundreds of organisms including plants and mammals. In most of the cases unidirectionally cloned cDNA libraries have been prepared using bacterial or phage vectors, so that the 5' and/or 3' end of the clones can be sequenced. Since single pass reads result in average ~350 high quality nucleotides, sequencing 3' ends covers mainly the untranslated region of the transcript. Moreover, the 3'end of the cDNA clones contain a long poly-A tail that is useless in terms of biological information and in general introduces technical difficulties in the sequencing process. However, because the untranslated 3'end represent the less conserved region of the transcripts it is useful, for example, to avoid misassembly of reads coming from highly conserved sequences from members of gene families. Sequencing 5' ends of unidirectional cDNA clones, on the other hand, can be of great advantage for large scale EST projects. Since the 5' untranslated region is shorter, it is likely that it contains protein-coding sequences. In addition, because a large proportion of clones present partial cDNA sequences, it is possible to collect enough information to assemble the full consensus sequence of a transcript, increasing the likelihood that database searches will result in the assignment of their putative functions. Based on this assumptions we decide sequence the 5' end of the cDNA clones to build up the SUCEST database.

The libraries

Table I shows the description of the libraries used in the SUCEST project. A variety of tissues were sampled from different cultivars, in order to access transcript information of genes expressed in many biological systems. Two libraries AD1 and HR1 were constructed using tissues from *in vitro* cultured plantlets infected with *Gluconacetobacter diazotroficans* and *Herbaspirilum diazotroficans*. These are endophytic nitrogen fixing bacteria that colonize sugarcane tissues (Lee, *et al.*, 2000). Sequencing from these libraries could lead to discovery of genes involved in plant-bacteria interaction and in nitrogen assimilation in sugarcane. Libraries AM1, AM2, LB1 and LB2 were constructed using apical meristem of young plants and lateral buds from adult plants. These libraries shall contribute with genes expressed at the initial stages of organ differentiation. Calli produced from sugarcane meristems was used in an experiment devised to access genes induced by cold and heat. Two weeks old calli was incubated at 4 °C or 37 °C for 12 h. Part of the tissues was maintained in the dark and part in continuous light. The CL6 library was prepared with a mixture of equal amounts of RNA extracted from these tissues and it is expected that this library will contribute with genes induced by cold and heat. FL1, FL3, FL4, FL5 and FL8 are libraries constructed from flower tissues harvested at different developmental stages and may contribute with genes expressed in this important plant organ. To access information on genes expressed in leaves, we constructed LR1 and LR2 libraries from leaf roll of adult plants and LV1 from etiolated leaves of plantlets grown *in vitro*. A collection of libraries representing roots or tissues from which roots emerge are represented by RT1, RT2 and RT3 which are libraries constructed from roots

sampled from plantlets grown *in vitro* or plants grown in greenhouse, while RZ1, RZ2 and RZ3 were constructed from root to shoot zone of young plants grown in greenhouse. SB1 is a library constructed from stalk bark of adult plants and may contribute with genes involved in the synthesis of cell wall components including waxes. SD1 and SD2 are libraries constructed from developing seeds. Finally, we constructed the libraries ST1 and ST3 from first and fourth internodes of adult plants at the time of intense sucrose synthesis and accumulation.

Quality control

Large-scale sequencing demands care with the quality of biological materials and accurate performance at each step of the process, both to provide sequence data of the highest possible quality and to detect or avoid mistakes (Adams *et al.*, 1995). At each step of the SUCEST project, from tissues sampling to sequence analysis, quality control and evaluation procedures were used to assess the accuracy of the data. The goal of the SUCEST project was that cDNA libraries should contain all sequences present in the initial poly(A)$^+$ mRNA population, which is useful to access expression profile through electronic Northern; unidirectionally cloned so that the orientation of each cDNA is known, facilitating subsequent sequence analysis; include a large proportion of full-length inserts; and reveal low levels of contamination with genomic or ribosomal RNA. Table II shows the quality control steps used during cDNA library construction and sequencing. Tissues were quickly frozen in liquid nitrogen, RNA quality analyzed by different methods and the cDNAs were synthesized and size selected using special gel filtration columns. cDNAs were unidirectionally cloned in pSPORT plasmidial vector and introduced into DH10B competent cells. Libraries with title less than 1 x 10^4 were discarded. Colonies were placed into 96 well plates and stored at -80 °C. A sample of ~400 clones from each library was examined to evaluate library quality, such as percentage of clones with no inserts, percentage of ESTs with exact matches to sequences derived from ribosomal RNA species, *E. coli* or bacteriophage lambda, percentage of ESTs with no significant matches to any sequence in the public databases, and an estimate of the number of clusters that contain a full-length coding region sequence. Libraries selected for EST analysis typically exhibited a broad diversity of transcripts (no single gene or small group of genes dominating the distribution), a low percentage of clones with no insert, a low percentage of ribosomal RNA clones, and no evidence of contamination with sequences from other organisms. The libraries that did not meet these general criteria were discarded.

Sequencing in the SUCEST project was carried out using ABI377 sequencers, which are prone to error during gel tracking. To minimize errors the 8th row of each 96 well plates was used to build control plates that were resequenced. Computer analysis was then used to check the address match. These allowed the SUCEST project to keep the address error to less than 5%, so that a sequence in the computer corresponds, with high fidelity, to a clone in the freezer.

Table II - Quality control and evaluation of SUCEST libraries.

Parameter	Quality control and evaluation
Tissue sampling	Tissues snap frozen quickly after harvesting
Poly(A)$^+$ RNA purification	Purity and RNA integrity were assessed by absorbance at 260/280 nm and agarose gel electrophoresis
cDNA synthesis	Tracer levels of ^{32}P used; agarose gel examination for degradation; column chromatography for size selection
cDNA library construction	Blue/white screen for inserts; PCR to check insert sizes; libraries must contain at least 10^5 recombinants
Library storage	All clones were grown in 96 well plates containing CG media supplemented with 8% glycerol > Plates were stored at -80 °C in triplicate
Sample sequencing	Around 400 clones of each library were sequenced to check gene diversity, contaminations and rRNA
Clone address	One clone in each twelve was resequenced to detect putative address mistakes
Template preparation	DNA quality and concentration checked by agarose gels

Quality control procedures for each step in the EST process are listed with specific points of evaluation or standards to be met.

SUCEST data set

Table III shows the summary of the complete data set of the SUCEST project. A total of 259,325 cDNA clones were sequenced in their 5'end region and 32,364 of them had also their 3'end region sequenced. Therefore, the project produced 291,689 ESTs. After trimming of low-quality sequences and removal of vector and ribosomal RNA ESTs, 237,954 ESTs potentially derived from protein-encoding messenger RNA (mRNA) remained. This represents a success index of 81.56%, which is comparable with other EST projects worldwide. Before entering the sequencing pipeline, each SUCEST cDNA library was evaluated for the average size of cDNA inserts. cDNA libraries that contained an average insert size below 500bp were discarded. The average insert size in all libraries was estimated to be 1,250bp (n = 4,000) (Table III). The distribution of the insert length was between 500 and 5,000bp. In order to clone genes encoding low molecular weight proteins, we constructed some cDNA libraries (LR2, RZ2 and SD2 - See Table IV) with an average insert size of 855bp.

Table III - Summary of SUCEST data.

Analyzed data	
Total ESTs	291,689
5' ESTs	259,325
3' ESTs	32,364
ESTs remaining after trimming quality control	237,954
Average insert size, bp	1,250
Average EST length, bp	750
Average EST bases with Phred quality ≥ 20	365

Numbers of sequenced cDNA clones and generated ESTs from 26 libraries constructed from different sugarcane tissues. 259,325 ESTs were generated by sequencing the 5' end of cDNA clones. Another 32,364 ESTs were generated by sequencing the 3' end of cDNAs clones. The average insert size was calculated for 400 cDNA clones from each library. The EST length and the number of bases with Phred quality ≥ 20 was calculated from the total EST set.

Table IV - Characteristics of the SUCEST libraries.

Library code	Average insert size (bp)	Sequenced clones	Valid reads	Success index (%)	Novelty (%)
AD1	1,330	18,144	14,701	81.02	55.34
AM1	1,300	12,480	10,881	87.19	55.05
AM2	-	15,648	13,403	85.65	49.45
CL6	1,150	7,392	5,518	74.65	63.62
FL1	1,400	18,528	15,343	82.81	54.82
FL3	1,340	13,056	10,727	82.16	53.26
FL4	1,370	16,896	13,964	82.65	52.19
FL5	1,180	10,080	7,744	76.83	66.05
FL8	1,400	5,184	4,652	89.74	72.26
HR1	-	12,000	9,729	81.08	52.11
LB1	1,150	7,488	5,879	78.51	62.91
LB2	1,660	10,560	8,953	84.78	60.33
LR1	1,240	14,112	11,701	82.92	56.85
LR2	870	4,128	3,418	82.80	68.13
LV1	1,260	6,432	4,557	70.85	67.32
RT1	1,450	8,640	7,255	83.97	58.26
RT2	1,400	12,288	10,606	86.31	54.86
RT3	1,000	10,560	7,441	70.46	58.54
RZ1	1,290	3,168	2,831	89.36	71.07
RZ2	-	5,760	5,031	87.34	63.14
RZ3	-	15,168	12,862	84.80	50.75
SB1	-	16,320	13,189	80.81	56.16
SD1	1,240	11,040	8,601	77.91	51.84
SD2	840	10,368	8,505	82.03	48.19
ST1	1,050	8,448	6,933	82.07	62.87
ST3	1,350	12,000	8,939	74.49	50.55

The average insert size of each library was determined in a sample of 400 clones by gel electrophoresis of the clones digested with PvuII. Valid reads are defined as reads containing at least 140 bp with Phred quality ≥ 20. The success index is the number of valid reads in relation to the number of clones sequenced. The Novelty represents the probability of a new sequence to be founding in the library.

After the trimming process, all new sequences were compared to the previous sequences that had already been deposited in the SUCEST database. Every time that an EST was similar to a sequence that already existed in the database, both were grouped together in a cluster. As noted in Table V, the 237,954 valid sequences were assembled into 43,141clusters.

Each cluster consensus sequence was compared against the non-redundant nucleotide and peptide databases (GenBank) using the programs BLASTN and BLASTX. Sequences that did not match these databases were further compared against the dbEST. Using a blast *E*-Value threshold (Altschul *et al.*, 1997) equal to or below e-5, of the 43,141 SUCEST clusters, 26,525 (61.5%) had matches with an existing sequence in GenBank (Table V). Therefore, 16,616 (38.5%) of the SUCEST clusters could potentially represent new genes. These values are comparable to those found for ESTs sequences from other organisms (Hwang *et al.* 2000; Adams *et al.* 1992; Claverie 1996). Ascribing functions to those anonymous sequences has therefore become one of the major bottlenecks in plant and animal genomics.

Tissue and cellular differentiation depend on specific patterns of gene expression. Therefore, in large-scale EST sequencing, sampling many different tissues and in different physiological conditions increases the chance to pick up transcripts rare in one cell type but less rare in another. SUCEST database was built up with sequences derived from 26 libraries constructed from different tissues sampled at different developmental stages (Table I) and an average of 10,000 clones were sequenced from each library. Sequencing from many libraries resulted in a novelty ratio as good as the ratios found in other EST projects that used normalized libraries (Bonaldo *et al.*, 1996).

Around 53.2% of the SUCEST clusters were formed by ESTs expressed in at least two libraries. This suggests that these genes are being coordinately expressed in different tissues or that they are expressed in response to specific physiological conditions or developmental requirements. On the other hand, 46.8% of the clusters (Table VI - the sum of specific contributions) are formed by ESTs expressed in only one library. This suggests that these ESTs could correspond to genes expressed in a tissue/time fashion, varying in different tissue/physiological conditions. Nonetheless, these data should be analyzed taking into account that 16,338 (37.9%) are singletons, therefore representing rare transcripts. The uniformity in the amount of singletons in the different libraries (Table VI) strengthens the value of the approach adopted.

A global analysis of all SUCEST clusters indicated that around 33% contain cDNA clones with full-length in-

Table V - Statistics of EST clustering and contiging.

ESTs analyzed	237,954
Total clusters (C+S)	43,141
Clusters with at least 2 reads (C)	26,303
Singletons (S)	16,838
C+S sequences finding homolog in GenBank	26,525
C+S sequences with no homolog in GenBank	16,616
C+S with full length insert	14,409

ESTs were clustered using CAP3 assembler (Huang and Madan, 1999). The *E*-value cut of threshold to be considered for C or S as having homology to other proteins in the nr GenBank database using BLASTX was ($<10^{-5}$). Clones were considered as having a putative full length insert when their sequences started within the first 15 amino acids of their hit in GenBank. C or S were considered as having tentative full consensus sequence when their sequences started within the first 15 amino acids and finished within the last 15 amino acid of their hit in GenBank.

Table VI - EST clustering in the individual libraries.

Library code	Number of clusters	Unique clusters	Number of singletons	Specific contribution (%)
AD1	10,736	3,120	2,821	3.84
AM1	7,870	1,930	1,726	2.37
AM2	9,079	2,389	2,012	2.94
CL6	4,282	1,231	1,112	1.51
FL1	11,438	3,740	3,468	4.60
FL3	7,847	2,178	1,997	2.68
FL4	10,145	2,626	2,407	3.23
FL5	6,489	1,697	1,589	2.08
FL8	3,963	811	780	0.99
HR1	6,697	1,664	1,434	2.04
LB1	4,697	1,149	1,074	1.41
LB2	7,056	1,749	1,597	2.15
LR1	8,867	2,250	2,104	2.77
LR2	2,901	696	662	0.85
LV1	4,005	1,037	950	1.27
RT1	5,706	1,435	1,336	1.76
RT2	7,851	2,081	1,875	2.56
RT3	5,699	1,398	1,251	1.72
RZ1	2,374	448	426	0.55
RZ2	4,054	939	869	1.15
RZ3	8,858	2,331	2,094	2.86
SB1	10,204	2,910	2,774	3.58
SD1	6,114	1,600	1,451	1.96
SD2	5,696	1,856	1,539	2.28
ST1	5,682	1,431	1,341	1.76
ST3	6,124	1,335	1,253	1.64

The number of clusters that contain one or more reads from a specific library is indicated, as well as, the clusters that were formed only by reads of a specific library (Unique Clusters). The number of clusters that were formed by only one read (Singleton) is also indicated. The specific contribution is calculated dividing the Unique Clusters of each library by the total number of clusters (43,141).

serts (Table V). This is in accordance with the results obtained in the mouse EST project (Marra *et al.*, 1999).

This collection of 237,954 ESTs provides us with a preliminary view into the gene expression profile of sugarcane. The identification of genes involved in different cellular processes suggests that the generation of large-scale ESTs should provide valuable insights into the molecular mechanisms of plant function and development.

REFRENCES

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and **Venter, J.C.** (1992). Sequence identification of 2,375 human brain genes [see comments]. *Nature 355* (6361): 632-634.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B. and **Moreno, R.F.** (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science 252* (5013): 1651-1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D. and **White, O.** (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature 377* (6547 Suppl): 3-174.

Altschul, S.F., Madden, T.L., Schiffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25* (17): 3389-3402.

Andrews, J., Bouffard, G.G., Cheadle, C., Lü, J., Becker, K.G. and **Oliver, B.** (2000). Gene Discovery Using Computational and Microarray Analysis of Transcription in the *Drosophila melanogaster* Testis. *Genome Res 10*: 2030-2043.

Bonaldo, M.F., Lennon, G. and **Soares, M.B.** (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res 6* (9): 791-806.

Claverie, J.M. (1996). Exploring the vast territory of uncharted ESTs. In: *Genomes, molecular biology and drug discovery*. Academic Press, pp. 56-71.

Green, P. (1999). phrap.doc: http://bozeman.genome.washington.edu/phrap.docs/phrap.html

Grivet, L. and **Arruda, P.** (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology 5*: 122-127.

Huang, X. and **Madan, A.** (1999). CAP3: A DNA sequence assembly program. *Genome Res. 9*: 868-877.

Hwang, D.M., Dempsey, A.A., Lee, C.Y. and **Liew, C.C.** (2000). Identification of differentially expressed genes in cardiac hypertrophy by analysis of expressed sequence tags. *Genomics 66* (1): 1-14.

Hwang, D.M., Dempsey, A.A., Wang, R.X., Rezvani, M., Barrans, J.D., Dai, K.S., Wang, H.Y., Ma, H., Cukerman, E., Liu, Y.Q., Gu, J.R., Zhang, J.H., Tsui, S.K., Waye, M.M., Fung, K.P., Lee, C.Y. and **Liew, C.C.** (1997). A genome-based resource for molecular cardiovascular medicine: toward a compendium of cardiovascular genes. *Circulation 96* (12): 4146-4203.

Lee, S., Reth, A., Meletzus, D., Sevilla, M. and **Kennedy, C.** (2000) Characterization of a Major Cluster of nif, fix, and

Associated Genes in a Sugarcane Endophyte, *Gluconacetobacter diazotrophicus*. *Journal of Bacteriology 182*: 7088-7091.

Liew, C.C., Hwang, D.M., Fung, Y.W., Laurenssen, C., Cukerman, E., Tsui, S. and **Lee, C.Y.** (1994). A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc Natl Acad Sci U.S.A. 91* (22): 10645-10649.

Manning, K. (1991). Isolation of nucleic acids from plants by differential solvent precipitation. *Anal Biochem 195* (1): 45-50.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L., Cardenas, M., Chamberlain, A., Chappell, J., Clifton, S., Favello, A., Geisel, S., Gibbons, M., Harvey, N., Hill, F., Jackson, Y., Kohn, S., Lennon, G., Mardis, E., Martin, J. and **Waterston, R.** (1999). An encyclopedia of mouse genes. *Nat Genet 21* (2): 191-194.

Soares, M. and **Bonaldo, M.** (2000). Constructing and screening normalized cDNA libraries. In: *Genome analysis: a laboratory manual* (Birren, B., Green, E., Klapholz, S., Myers, R. and Roskams, A., eds.), CSHL Press: CSHL Press, pp. 49-157.

Simpson, A.J.G. and **Perez, J.F.** (1998). Latin America - ONSA, the Sao Paulo virtual genomics institute. *Nat Biotechnol 16*: 795-796.

Telles, G.P., Braga, M.D.V., Dias, Z., Lin, T., Quitzau, J.A.A., da Silva, F.R. and **Meidanis, J.** (2001). Bioinformatics of the sugarcane EST project. *Genetics and Molecular Biology 24* (1-4): 9-15.

Telles, G.P. and **da Silva, F.R.** (2002). Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology 24* (1-4): 17-23.

STIC-ILL

N05/23

44739

From: Zeman, Mary
Sent: Friday, May 23, 2003 11:26 AM
To: STIC-ILL
Subject: References 09/961058

Please send a copy of the following references
Thank you
Mary K. Zeman
Examiner, 1631
305-7133
CM1 12A17
mailbox: CM1 12D01

COLLECTED

SO Methods in Molecular Biology (Totowa, New Jersey) (2000), 132(Bioinformatics Methods and Protocols), 71.91
CODEN: MMBIED; ISSN: 1064.3745

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 1.7 ISSN: 1415.

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 9.15 ISSN: 1415.4757

SO GENETICS AND MOLECULAR BIOLOGY; (2001) 24, 1.4, 17.23 ISSN: 1415.4757

SO International Genome Sequencing and Analysis Conference, (2000) Vol. 12, pp. 97.98. print.
Meeting Info.: 12th International Genome Sequencing and Analysis Conference Miami Beach, Florida, USA September 12.15, 2000

Genetics and Molecular Biology, 24, (1-4), 17-23 (2001)

Trimming and clustering sugarcane ESTs

Guilherme P. Telles[1] and Felipe R. da Silva[2]

Abstract

The original clustering procedure adopted in the Sugarcane Expressed Sequence Tag project (SUCEST) had many problems, for instance too many clusters, the presence of ribosomal sequences, etc. We therefore redesigned the clustering procedure entirely, including a much more careful initial trimming of the reads. In this paper the new trimming and clustering strategies are described in detail and we give the new official figures for the project, 237,954 expressed sequence tags and 43,141 clusters.

INTRODUCTION

The Sugarcane EST project (SUCEST) produced 291,689 expressed sequence tags (ESTs) (Adams *et al.*, 1991). In the pipeline of the project it was important to cluster together sequences from the same transcript molecule and to obtain a representative sequence for each group. Clustering was important to evaluate the redundancy of the set of ESTs during library production and sequencing, and at the end of the project. Clustering also produces a smaller set of sequences which facilitates investigation of the data by biologists and computer scientists (Telles *et al.*, 2001).

As in any other EST project, the raw SUCEST sequences sometimes contained unwanted segments like polyadenylation (poly-A), regions with low base quality, fragments from vectors and adapters, and slippage. Some reads may also came from ribosomal RNA or contaminant DNA. Such segments are unwanted because they introduce similarity between ESTs that has no relevance for clustering, and removal of such segments is essential to cluster correctly.

Trimming and clustering procedures were established at the beginning of the SUCEST project in July 1999, but the amount of data grew each day and it soon became clear that the trimming and clustering procedures were both not good enough. SUCEST data-users were pointing out many problems when we designed and implemented new trimming and clustering procedures.

A trimming procedure is essentially the task of searching ESTs for unwanted regions, identifying them and then deciding whether to remove the unwanted region or to discard the entire EST. Trimming has already been described for UniGene (www.ncbi.nlm.nih.gov/UniGene), TIGR Gene Indices (Quackenbush *et al.*, 2000) and STACK (Miller *et al.*, 1999).

In the SUCEST project, clustering was always performed using a fragment assembler for the whole set of ESTs. This is different from the procedure used by Unigene, TIGR Gene Indices, JESAM (Parsons and Rodrigues-Tomé, 2000) and STACK which use some kind of pairwise comparison to estimate distance between ESTs, build clusters and then, if ever, assemble the clusters separately. In its first version, SUCEST clustering scheme produced 81,223 clusters (41,582 singletons) while the current version has 43,141 clusters (16,838 singletons).

In this paper we describe trimming in detail, because it had a major influence on the work performed by the assembler at the clustering stage. We have also compared the results of different assemblers for our set of ESTs before we decide in favor of the CAP3 program (Huang and Madan, 1999). Although we had confidence in the fragment assemblers comparison performed by Liang *et al.* (2000), three issues motivated us to produce our own comparison routines. Firstly, we wanted to examine the assembly results for our particular set of ESTs, secondly, we were using ESTs quality data and, thirdly, we used parameters for the assemblers that differ from the default ones. We also introduce the trimming and clustering procedures early in the project. Our intention in this paper is not to emphasize our improved results but to show the remarkable effect that 'noise' (*i.e.* unwanted sequences) can have on clustering.

METHODOLOGY AND RESULTS

Clone libraries were prepared as described by Vettore *et al.* (2001) and sequenced by ABI 377 (Applied Biosystems) machines. After being processed by the phred base-calling program (version 0.980904.e, www.phrap.org) and by the phd2fasta program (version 0.990622.d, www.phrap.org), ESTs were stored as fasta and quality files in the 5' to 3' orientation. These files contained 291,689 sequences with an average length of 864.5 ± 186.3

[1]*Bioinformatics Laboratory, Institute of Computing, UNICAMP, CP 6176, 13083-970 Campinas, SP, Brazil.*
[2]*Center for Molecular Biology and Genetic Engineering, UNICAMP, CP 6010, 13083-970 Campinas, SP, Brazil.*
Send correspondence to Guilherme P. Telles. E-mail: pimentel@ic.unicamp.br.

bases. The average number of bases with a phred quality value greater than 20 per read was 399.5 ± 151.3. The programs were run on an 8 GB RAM AlphaServer ES40 (Compaq) with 2 processors at 667 MHz executing the OSF1 operating system (version 4.0G).

Trimming

An EST set may contain unwanted sequences made up of poly-A fragments, vector and adapter fragments, low quality ends, ribosomal RNA, contaminant DNA and slipped sequences. When clustering the sequences to produce groups of transcripts, these unwanted sequence introduce irrelevant relationships between reads. Trimming is the removal of such regions from ESTs or the removal of entire ESTs from the set.

Trimming refined the reads in several steps, using the blastall program (version 10/31/2000, www.ncbi.nlm.nih.gov) that implements the BLAST algorithm (Altschul *et al*, 1997), the cross-match program (version 0.990319, www.phrap.org), the SWAT program (version 0.990319, www.phrap.org) and *ad hoc* pattern-matching programs written in Perl (version 5.6.0, www.cpan.org). Parsers (programs that do some kind of interpretation on data based on its syntactical structure) for the output of these programs were written in Perl, and bash (version 2.04.0(1), www.gnu.org) scripts were used to filter, build histograms and summarize data. Some regions, like poly-A, were searched several times, each time with a different recognition criterion. Trimming was tuned to keep as much as possible from each sequence.

The trimming scheme is summarized in Figure 1. The first step was the removal of ribosomal RNA sequences, and for this the ESTs were compared against 18S rRNA from *Zea mays* (GenBank AF168884), 5.8S rRNA from *Platanus occidentalis* (GenBank AF162215) and 26S rRNA from *Lambertia inermis* (GenBank AF274652) using the BLAST program. The choice of these rRNA sources was based on the similarity between them and sugarcane rRNA. A match with an e-value less than 10^{-10} was the threshold to discard a read, a total of 8,473 reads being removed in this step.

The next step was vector and adapter sequence masking, using the cross-match program that replaced bases with an X if they were very similar to vector and adapter sequences used in the clone libraries. This was followed by removing the vector and adapter sequences themselves from the reads by deleting the regions marked with an X. The actual treatment given to these ambiguous regions depended on where the X-regions were found and how many there were, an X-region being a contiguous masked sub-sequence in a read.

Classes were devised based on the analysis of histograms of the lengths of X-regions, distance of the X-region

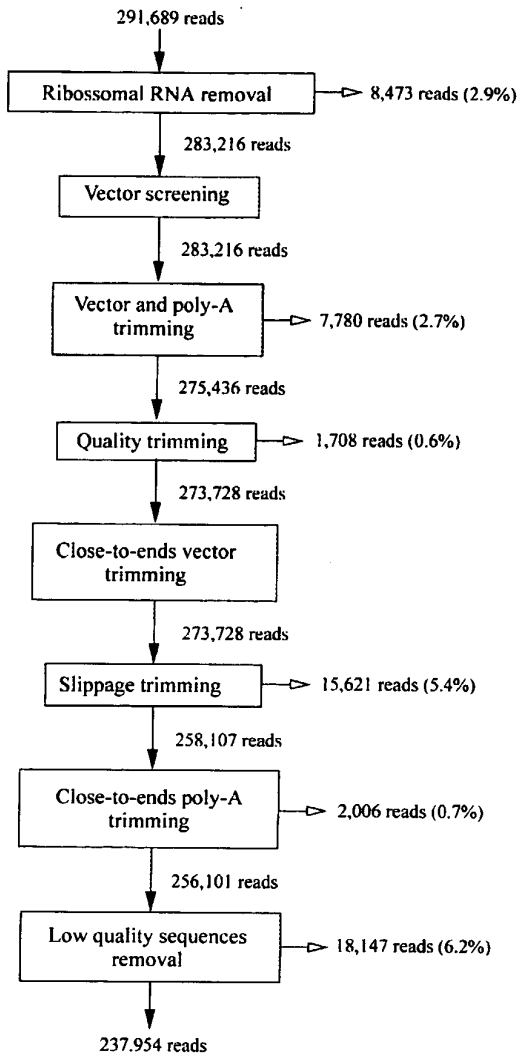

Figure 1 - Overview of trimming procedure. White-headed arrows indicate the number of reads discarded in each step, with the percentage of total shown in parenthesis.

from the 5' and 3' ends and on the analysis of the number of ESTs falling into each class. These classes were as follows:

Class 1. There were two distinct X-regions in the read, this being what is to be expected as the result of sequencing a clone with a small insert. In this case only the sequence between the X-regions was kept.

Class 2. There were more than two X-regions in the read, probably because of a low-quality vector. In this case we did not change the read.

Class 3. There was only one X-region of no more than 300 bases that was less than 50 bases away from the 5' end. This was the case when the region from the X-region down to the 5' end probably consisted of vector sequences extending from the sequencing priming site to the cloning

site. In this case we removed the X region together with the 5' end.

Class 4. There was only one X-region with more than 300 bases that was less than 50 bases away from the 5' end. In this case the clone probably had no insert so we discarded the whole read.

Class 5. There was only one X-region of at most 300 bases that was 51 to 300 bases from the 5' end. In this case it was hard to decide what the insert was so the read was not changed.

Class 6. There was only one X-region with more than 300 bases that was 51 to 300 bases from the 5' end. This probably occurred when the X-region and the 3' end consisted of a vector sequence after the cloning site. In this case we removed both the X-region and the 3' end.

Class 7. There was only one X-region of any length and it was at least 300 bases away from the 5' end. In this case we again removed both the X-region and 3' end because the deleted region probably consisted of a post cloning-site vector sequence.

While removing X-regions any poly-A fragment close to them was also removed. A poly-A fragment was considered to be any region that scored at least 8 when aligned with a probe sequence of As (adenines) only. The scoring scheme added 1 for a match and -2 for a mismatch, gaps were given a high penalty (-8) because they should not occur. The poly-A had to be at most 10 bases away from X-regions. Alignments were performed using the SWAT program. Depending on the reading direction a poly-A can be read as poly-T, so a poly-T probe was used as well. The removal of X-regions discarded 7,780 sequences.

The next step was quality-trimming, for which a window of 20 bases was slid over every sequence in the set. Starting at the 3' end, the window was slid one base at a time, dropping the extreme base until 12 or less bases in the window had a quality value below 10, the process being repeated for the 5' end. After quality-trimming, X-regions not further than 10 bases away from an end were removed. Quality-trimming removed 1,708 sequences from the set.

The quality-trimming thresholds were chosen as follows. A subset of 10,000 SUCEST sequences was randomly selected on the basis of (i) high similarity (BLASTX e-value below 10^{-20}) with protein sequences in the NCBI nr database (www.ncbi.nlm.nih.gov), (ii) the length of the matching nr sequence was enough to cover the EST and (iii) the region of similarity did not extend to the end of the EST. By using these criteria we had matches showing a region of similarity that could, potentially, extend to the end of an EST. Cases where the region of similarity did not extend to the end of the EST may have been due to the low quality of the EST sequence.

The exact point where the region of similarity ended, the 'BLAST hit end' (BHE), was recorded for each EST in the set and then the set went through the quality-trimming procedure with varying values for the length of the window, quality threshold and number of bases below threshold. Obviously, high quality thresholds and low numbers of bases produced shorter reads. The difference between the trimmed position (TP) and the BHE (TP-BHE) was calculated and averaged. The results for a 20-base quality window are shown in Figure 2. The square in the figure indicates the selected threshold values and shows that, on average, 43 bases after the BLAST hit end were kept.

The next step was slippage-trimming, slippage being a sequencing artifact (Anon, 1998) which produces 'echoed' bases in sequences, *i.e.* for one occurrence of a nucleotide in the template the chromatogram shows several peaks (q.v. Figure 3). Although bases sometimes appeared with high 'background noise' (*e.g.* bases 215-230), generally the intensity of the echoed peak was such that the base caller incorrectly assigned a high quality value for the fake bases (*e.g.* bases 175-205) and this prevented quality-trimming of these artifacts.

A method to identify slipped reads based on the sequence of the read was devised, this method being able to find reads having many regions with repetitive bases (echoed regions). The product of echoed regions lengths (with at least 5 bases) was evaluated for each sequence. Echoed regions larger than 10 bases contributed 10 to the product only. Sequences with a product greater than 10^8 and echoed regions covering more than 20% of its length were discarded completely. This was the procedure adopted in most cases when slippage was caused by a long poly-A sequence at the 5' end of the read. But when a long poly-A at the 3' end was increasing the product only the poly-A (together with the remaining 3' sequence) was discarded. The threshold for poly-A identification in this situation was an alignment with a score of at least 160. These thresholds were determined by varying the parameters for echoed region recognition, evaluating the products, and looking at several chromatograms in many product ranges. Slippage-trimming removed 15,621 reads.

The next step in the trimming procedure was another poly-A/T removal round, where poly-A/T scoring at 280 and over was removed from sequences. Smaller poly-A/T,

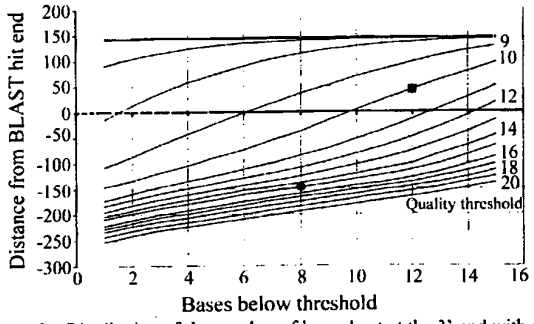

Figure 2 - Distribution of the number of bases kept at the 3' end with a quality window size of 20, with respect to the best BLAST hits against nr (see text). The square and the bullet indicate the values used in the new and old trimming procedures, respectively.

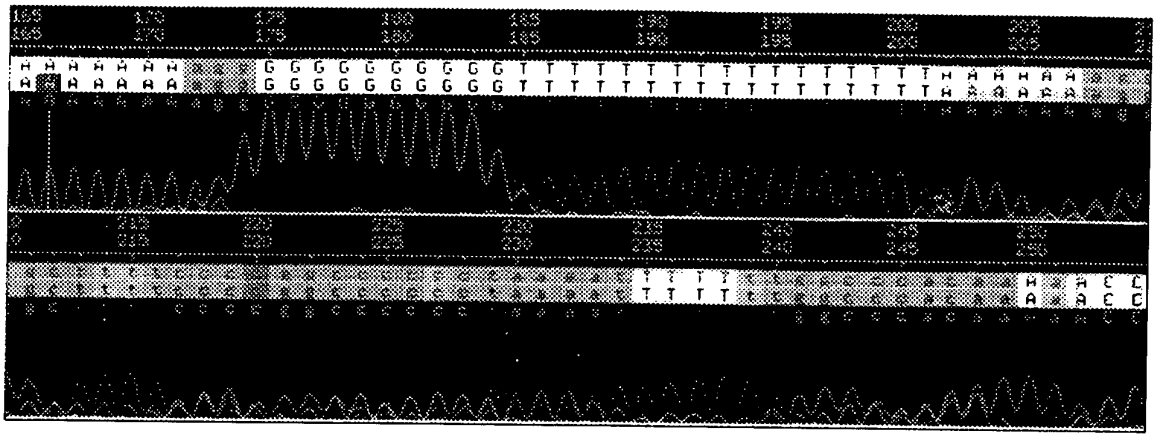

Figure 3 - Consed (www.phrap.org) trace window of a slipped read. The background of the base letters indicates their phred quality. Darker colors correspond to lower qualities. The numbers above the letters show their position in the read.

scoring at least 30 and less than 20 bases away from one of the ends was also removed. This step removed 2,006 reads.

The final step was to remove any read with less than 100 bases or with less than 50 bases having phred quality greater or equal to 20. A total of 18,147 reads fell in this case.

At the end of all the steps described above, 237,954 reads were left with an average length of 641.6 ± 139.8 bases (152.5 Mbp in total). The average number of bases with a phred quality greater or equal to 20 per read was 397.8 ± 120.1.

In contrast, the trimming method formerly used in the SUCEST project was simpler. That method started with only one round of very restricted poly-A removal, searching for 12 or more consecutive As adjacent to the vector. The final step was quality-trimming using the same scheme as above with a window length of 20, quality equal to 15 and the number of bases equal to 8. For the reads used in the quality window experiment this combination of thresholds discarded 137 bases from the reads on average (relative to the BHE) as shown in Figure 2. This method applied to the original set of SUCEST reads resulted in 261,609 reads with average length of 512.1 ± 114.8 bases. The average number of bases with a phred quality greater or equal to 20 per read was 392.4 ± 128.3.

BLAST was used to compare the ESTs from the original set of reads in the SUCEST database with the genomes of *Xylella fastidiosa*, *Xanthomonas citri*, *Escherichia coli* and other potential laboratory contaminants that could have been present in the libraries. A match of at least 100 bases and more than 90% identity resulted in the read being marked as probably resulting from contamination. A total of 114 ESTs were thus marked. Because there were so few matches, and the difficulty of deciding whether or not marked ESTs really were the result of contamination, these ESTs were not removed by either of the trimming procedures.

Clustering

For the SUCEST project it was necessary to estimate the redundancy of the clone libraries as they were sequenced, which could be achieved by joining similar transcripts into clusters. Clustering results allowed project coordinators to decide when to stop sequencing any particular library.

Fragment assemblers were used for clustering. A fragment assembler is a program that takes a set of reads and their qualities as input, builds groups based on the overlaps of reads and creates a consensus sequence for the reads in each group.

Reads processed by the old trimmer were assembled using the phrap program(version 0.990319, www.phrap.org) with the arguments set to predetermined values (penalty -15, bandwidth 14, minscore 100, shatter_greedy) which made it more stringent and with quality data. This assembly, called 'old-trim', produced 81,223 clusters (41,582 singletons).

To cluster the reads trimmed by the new procedure, three different assemblies were performed and compared. Phrap was used with two sets of arguments, the default arguments (phrap-d assembly) and the more stringent arguments listed above (phrap-hs assembly). The CAP3 program was used with its default arguments. Quality data was used for every assembly. Table 1 shows the cluster size (number of ESTs in a cluster) distribution for the assemblies, as well as the number of equal clusters between them. Equal clusters are those with the same reads.

Two tests were performed for the assemblies. The first verified 'internal consistency' by checking every cluster with two or more reads for discrepant reads. To be

Table 1 - Cluster sizes distribution for CAP3, phrap-d and phrap-hs assemblies by the new trimming procedure. The 'X' columns indicate the number of equal clusters between two assemblies, while the 'common' column shows the number of clusters equal in the three assemblies. The number of clusters obtained with the original trimming procedure are shown in the 'Old-trim' column. Cluster sizes represent the number of expressed sequence tags (ESTs) in a cluster.

Cluster size	phrap-hs	X	phrap-d	X	CAP3	X phrap-hs	Common	Old-trim
1	32202	13731	18535	11634	16838	14296	10744	41582
2	12440	5617	9207	4869	7665	4852	3792	13619
3	6752	2402	5192	2151	4193	1984	1441	7421
4	4225	1239	3329	1145	2709	992	697	4482
5	2856	676	2360	700	1872	521	344	3110
6	2098	442	1806	482	1452	354	231	2151
7	1582	288	1362	317	1115	220	144	1582
8	1245	202	1091	242	862	153	99	1219
9	974	156	913	186	720	113	72	964
10	776	105	752	143	634	74	44	809
11	639	76	607	99	511	54	30	641
12	492	71	547	99	429	46	32	490
13	437	47	454	90	400	40	25	430
14	366	42	391	40	341	26	13	366
15	306	31	390	50	295	18	11	312
16	273	25	279	35	275	18	8	257
17	225	15	273	23	235	11	4	206
18	177	11	227	15	191	5	2	183
19	124	6	177	18	176	5	3	153
20	143	10	149	13	179	6	3	136
21	113	6	130	5	133	2	0	106
22	105	3	130	5	117	2	1	98
23	92	4	100	9	140	3	2	79
24	80	4	99	6	122	2	1	82
25	69	3	109	6	86	5	2	60
26	56	2	108	9	72	1	1	59
27	51	2	59	4	78	1	1	49
28	44	1	73	5	74	1	1	46
28	439	5	857	25	1227	0	0	577
Total	69381	25222	49706	22425	43141	23805	17748	81223

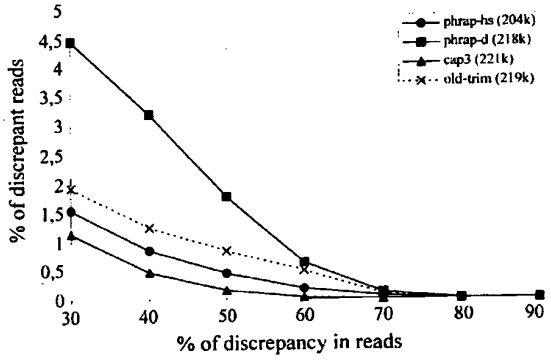

Figure 4 - Distribution of discrepant reads among the assemblies. As discrepant reads can only be calculated for clusters of two or more reads the number of reads belonging to such clusters in each assembly is shown in parentheses in the legend.

discrepant, a read base must both disagree with the consensus base and have less than a 2% probability of being miscalled by the phred program. An x% discrepant read is a read with at least x% discrepant bases. Figure 4 shows the proportion of x% discrepant reads in each assembly, for values of x varying from 30 to 90 in steps of 10.

The second test verified the 'external consistency' of the assemblies by comparing the consensi produced by a given assembly to each other using BLAST. Percentage identity was evaluated for end-overlaps of 200 or more bases found between two clusters, and Figure 5 shows a plot of the percentage of clusters having an identity of more than 75% with other clusters in a given assembly with respect to the total of possible overlaps within that set of clusters.

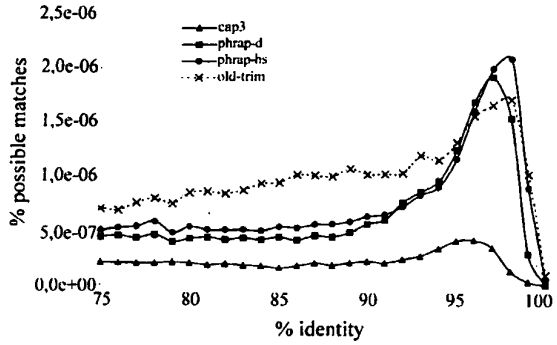

Figure 5 - Plot of external consistency test results. For a given assembly with n clusters the number of overlaps detected was divided by n(n-1)/2, which is the maximum number of possible overlaps for n clusters.

DISCUSSION

The trimming procedure described in this paper discarded 53,735 SUCEST reads, 18.4% of the total. In spite of this large number, it is worth noting that 16% of the discarded sequences were ribosomal RNA and 34% were smaller than 100 bases. We cannot exclude the possibility of that we have discarded useful reads with this procedure, but we tried to avoid this as much as possible. It is also obvious that not every artifact has been removed. For example, counting how many reads have a sub-sequence of at least 30 consecutive adenines in the output of the trimming procedure found 711 reads. Moreover, trimming is not a light computational task, taking 8.3 h to process all the SUCEST reads.

Nevertheless, the influence of the quality of trimming on the final clustering is remarkable. For instance, it is hard to accept that the number of singletons in the old assembly are uniquely expressed sugarcane genes, and 81,223 was an unreasonably large number of clusters. Good trimming also shortened the CPU time required for clustering, the phrap program took 9.2 h to build the phrap-d assembly and 6.5 h for phrap-hs assembly, while the CAP3 program took 77.1 h. To assemble the old set of trimmed reads, phrap took 5 times more time than it spent to produce phrap-hs, while CAP3 ended abnormally when fed with that data-set.

We have used a fragment assembler for the whole set of ESTs in the SUCEST database and, consequently, the biological definition of 'one cluster, one gene' cannot be used. A SUCEST cluster can be better defined as 'a set of very similar transcripts'.

Building consensus sequences for clusters is useful in several respects. Firstly, electing a representative sequence for each cluster results in a smaller set of sequences to work with. Secondly, the portions of representative sequences covered by more than one read are more accurate than the reads themselves. Thirdly, representative sequences may be longer than individual reads, increasing their usefulness. This third point was confirmed by the fact that 33% of representative sequences with homologous genes in other organisms were actually full-length sequences (Vettore *et al.*, 2001).

However, chimeras may result from assembling ESTs and a further problem is that using a fragment assembler for clustering will put alternatively spliced forms of genes into different clusters. But in a dodecaploid organism like sugarcane it is especially difficult to distinguish alleles of genes from very conserved multigene families based on similarity.

The assembly produced by the CAP3 program was taken as the 'official' clustering for the SUCEST project. This decision was based on the result of the internal and external consistency tests, where the CAP3 assembly outperformed both the phrap-hs and phrap-d assemblies. Internal consistency shows that the CAP3 assembly has a lower incidence of discrepant reads in clusters when compared to the other assemblies. External consistency reveals that the CAP3 program produces fewer redundant clusters, *i.e.* two or more clusters that probably should be condensed to a single cluster. Unfortunately, we performed no comparisons of our results with those that would be produced using some other method described in the literature. This is an interesting investigation to perform in the future.

The trimming and clustering procedures described in this paper hide a large amount of computational time and human work spent looking at the data, testing insights, adjusting parameters, and designing the pipeline. There are no 'magic numbers'. We believe that these guidelines may be used in some other EST projects, although using these procedures with different data sets may require some adjustments. The need for many cycles of adjustment and testing is a natural consequence of the nature of the noise present in ESTs, the limitations posed by technological issues and the lack of a complete understanding of the biological processes occurring within cells.

RESUMO

O método de *clustering* adotado no Projeto SUCEST (Sugarcane EST Project) tinha vários problemas (muitos *clusters*, presença de seqüências de ribossomo etc.) Nós assumimos a tarefa de reprojetar todo o processo de *clustering*, propondo uma "limpeza" inicial mais cuidadosa das seqüências. Neste artigo as estratégias de limpeza das seqüências e de *clustering* são descritas em detalhe, incluindo os números oficiais do projeto (237,954 ESTs e 43,141 *clusters*).

ACKNOWLEDGMENTS

This work was supported by the Brazilian agencies FAPESP and CNPq.

REFERENCES

Anon. (1998). *Chemistry Guide for Automated DNA Sequencing*, Applied Biosystems, Foster City-CA, USA, 242 pp.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R., Kerlavage, A.R., McCombie, W.R. and **Venter, J.C.** (1991). Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science 252*: 1651-1656.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*: 3389-3402.

Huang, X. and **Madan, A.** (1999). CAP3: a DNA sequence assembly program. *Genome Res. 9*: 868-877.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. and **Quackenbush, J.** (2000). An optimized protocol for analysis of EST sequences. *Nucleic Acids Res. 28* (18): 3657-3665.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J.A., Ptitsyn, A.A., Broveak, T.R. and **Hide, W.A.** (1999). A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res. 9*: 1143-1155.

Parsons, J. and **Rodriguez-Tomé, P.** (2000). JESAM: CORBA software components to create and publish EST alignments and clusters. *Bioinformatics. 4* (16): 313-325.

Quackenbush, J., Liang, F., Holt, I., Pertea, G. and **Upton, J.** (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res. 28* (1): 141-145.

Telles, G.P., Braga, M.D.V., Dias, Z., Lin, T.-L., Quitzau, J.A.A., da Silva, F.R. and **Meidanis, J.** (2001). Bioinformatics of the Sugarcane EST Project. *Genetics and Molecular Biology 24* (1-4): 9-15.

Vettore, A., da Silva, F.R., Kemper, E. and **Arruda, P.** (2001). The libraries that made SUCEST. *Genetics and Molecular Biology 24* (1-4): 1-7.

L Number	Hits	Search Text	DB	Time stamp
1	111	sequence and database and (trim or trimming or trimmed) and input and search and directory and spreadsheet	USPAT; US-PGPUB	2003/05/23 10:42
2	0	sequence and database and (trim or trimming or trimmed) and input and search and directory and spreadsheet and (DNA or nucleic or RNA or polynucleotide or oligonucleotide)	USPAT; US-PGPUB	2003/05/23 10:42
3	0	database and (trim or trimming or trimmed) and input and search and directory and spreadsheet and (DNA or nucleic or RNA or polynucleotide or oligonucleotide)	USPAT; US-PGPUB	2003/05/23 10:43
4	0	database and (trim or trimming or trimmed) and directory and spreadsheet and (DNA or nucleic or RNA or polynucleotide or oligonucleotide)	USPAT; US-PGPUB	2003/05/23 10:43
5	13	database and (trim or trimming or trimmed) and spreadsheet and (DNA or nucleic or RNA or polynucleotide or oligonucleotide)	USPAT; US-PGPUB	2003/05/23 10:47
7	68	database and ((trim or trimming or trimmed) adj50 (DNA or nucleic or RNA or polynucleotide or oligonucleotide))	USPAT; US-PGPUB	2003/05/23 11:03
6	1469	database and (trim or trimming or trimmed) and (DNA or nucleic or RNA or polynucleotide or oligonucleotide)	USPAT; US-PGPUB	2003/05/23 11:03
8	1	database and ((trim or trimming or trimmed) and (DNA or nucleic or RNA or polynucleotide or oligonucleotide))	EPO; JPO; DERWENT; IBM_TDB	2003/05/23 11:04

```
      INDEX 'ADISCTI, ADISINSIGHT, ADISNEWS, AGRICOLA, ANABSTR, AQUASCI,
      BIOBUSINESS, BIOCOMMERCE, BIOSIS, BIOTECHABS, BIOTECHDS, BIOTECHNO, CABA,
      CANCERLIT, CAPLUS, CEABA-VTB, CEN, CIN, CONFSCI, CROPB, CROPU, DDFB,
      DDFU, DGENE, DRUGB, DRUGLAUNCH, DRUGMONOG2, ...' ENTERED AT 11:09:11 ON
      23 MAY 2003
                SEA SEQUENCE AND DATABASE AND TRIM? AND (COMPUTER OR SOFTWARE O
               ---------
                 6   FILE BIOSIS
                12   FILE BIOTECHABS
                12   FILE BIOTECHDS
                 5   FILE BIOTECHNO
                 1   FILE CABA
                 1   FILE CANCERLIT
                 9   FILE CAPLUS
                 1   FILE CEABA-VTB
                 8   FILE CEN
                 1   FILE DDFU
                11   FILE DGENE
                 2   FILE DRUGU
                 1   FILE EMBAL
                 7   FILE EMBASE
                 4   FILE ESBIOBASE
                 4   FILE FEDRIP
                 6   FILE GENBANK
                44   FILE IFIPAT
                 2   FILE LIFESCI
                 8   FILE MEDLINE
                 1   FILE PASCAL
                 1   FILE PHARMAML
                 1   FILE PHIN
                94   FILE PROMT
                 1   FILE RDISCLOSURE
                 3   FILE SCISEARCH
                 1   FILE TOXCENTER
              4334   FILE USPATFULL
                87   FILE USPAT2
                 4   FILE WPIDS
                 4   FILE WPINDEX
L1               QUE SEQUENCE AND DATABASE AND TRIM? AND (COMPUTER OR SOFTWARE O
                 ---------

      FILE 'BIOTECHDS, CAPLUS, CEN, MEDLINE, EMBASE, BIOSIS, BIOTECHNO,
      ESBIOBASE, FEDRIP, WPIDS, SCISEARCH, DRUGU, LIFESCI, CABA, CANCERLIT,
      CEABA-VTB, EMBAL, PASCAL, PHARMAML, PHIN, RDISCLOSURE, TOXCENTER' ENTERED
      AT 11:12:50 ON 23 MAY 2003
L2             83 S L1
L3             49 DUP REM L2 (34 DUPLICATES REMOVED)
L4             49 FOCUS L3 1-


COST IN U.S. DOLLARS                                SINCE FILE      TOTAL
                                                      ENTRY    SESSION
FULL ESTIMATED COST                                  121.64     125.15

DISCOUNT AMOUNTS (FOR QUALIFYING ACCOUNTS)          SINCE FILE      TOTAL
                                                      ENTRY    SESSION
CA SUBSCRIBER PRICE                                   -5.21      -5.21

STN INTERNATIONAL LOGOFF AT 11:17:41 ON 23 MAY 2003
```